

Análise de dados do serviço Via Ipê:

Caracterização e predição de falhas

Vitor Zanotelli, Giovanni Comarela, Rodolfo Villaça, Magnos Martinello



Problema de Pesquisa

É possível criar um sistema/modelo capaz de prever falhas em enlaces de rede?

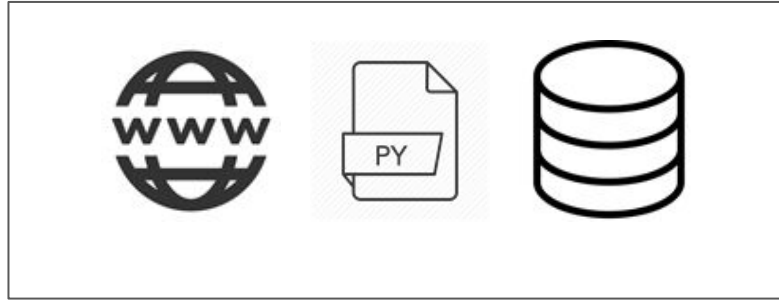
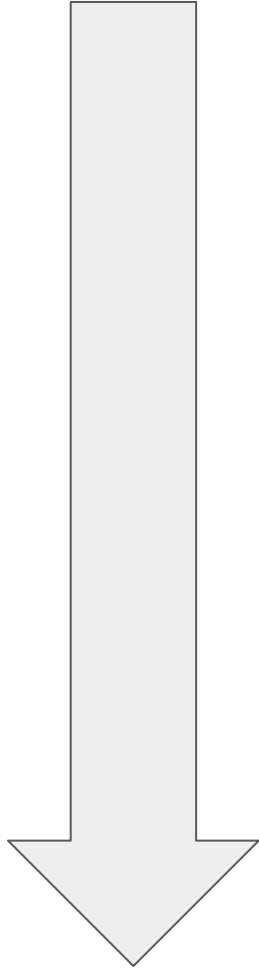


Via Ipe

<https://viaipe.rnp.br>

Disponibilização de métricas do funcionamento da rede Ipê.





Web Scraping



Análise Exploratória de Dados



Criação de Modelos de
Aprendizado de Máquina

Coleta de Dados

O serviço Via Ipê disponibiliza arquivos JSON contendo os dados da rede Ipê minuto a minuto através do link: <https://viaipe.rnp.br/data/>

Início da data de coleta para realização dos estudos: Março de 2020

Descrição dos Dados

Dentre os valores presentes no JSON fornecido, podemos destacar as seguintes variáveis de interesse:

- **Tempo:** DateTime com precisão de minuto (Y-M-D h:m:s).
- **Interface:** String com o identificador da interface na rede.
- **Local:** String com a abreviação do Estado.
- **Latitude e Longitude:** Valores da localização geográfica por interface.
- **Taxa de Download:** Medida em bps (bits por segundo).
- **Taxa de Upload:** Medida em bps (bits por segundo).
- **RTT:** Medido em ms (microsegundos).
- **Packet Loss:** Percentual de perda de pacotes (0 a 100%) na interface.

Seleção do Conjunto de Dados

Para um trabalho inicial, selecionamos duas fatias de tempo totalizando 14 dias:

No mês de abril do dia 16 ao 22.

- Volume de dados : [17.195.298 linhas x 12 colunas]
- Cada linha é uma interface

No mês de julho do dia 10 ao 16

- Volume de dados : [14.357.092 linhas x 12 colunas]

Dentre as variáveis de interesse selecionamos:

Interface, Tempo, Local, e Packet Loss.

Classificação do Packet Loss

A Via Ipê classifica a qualidade do sinal numa interface de acordo com o percentual de perda de pacotes da seguinte forma:

- Abaixo de 0.01%: Ótima
- Entre 0.01% e 1%: Boa
- Entre 1% e 3%: Regular
- Acima de 3%: Ruim

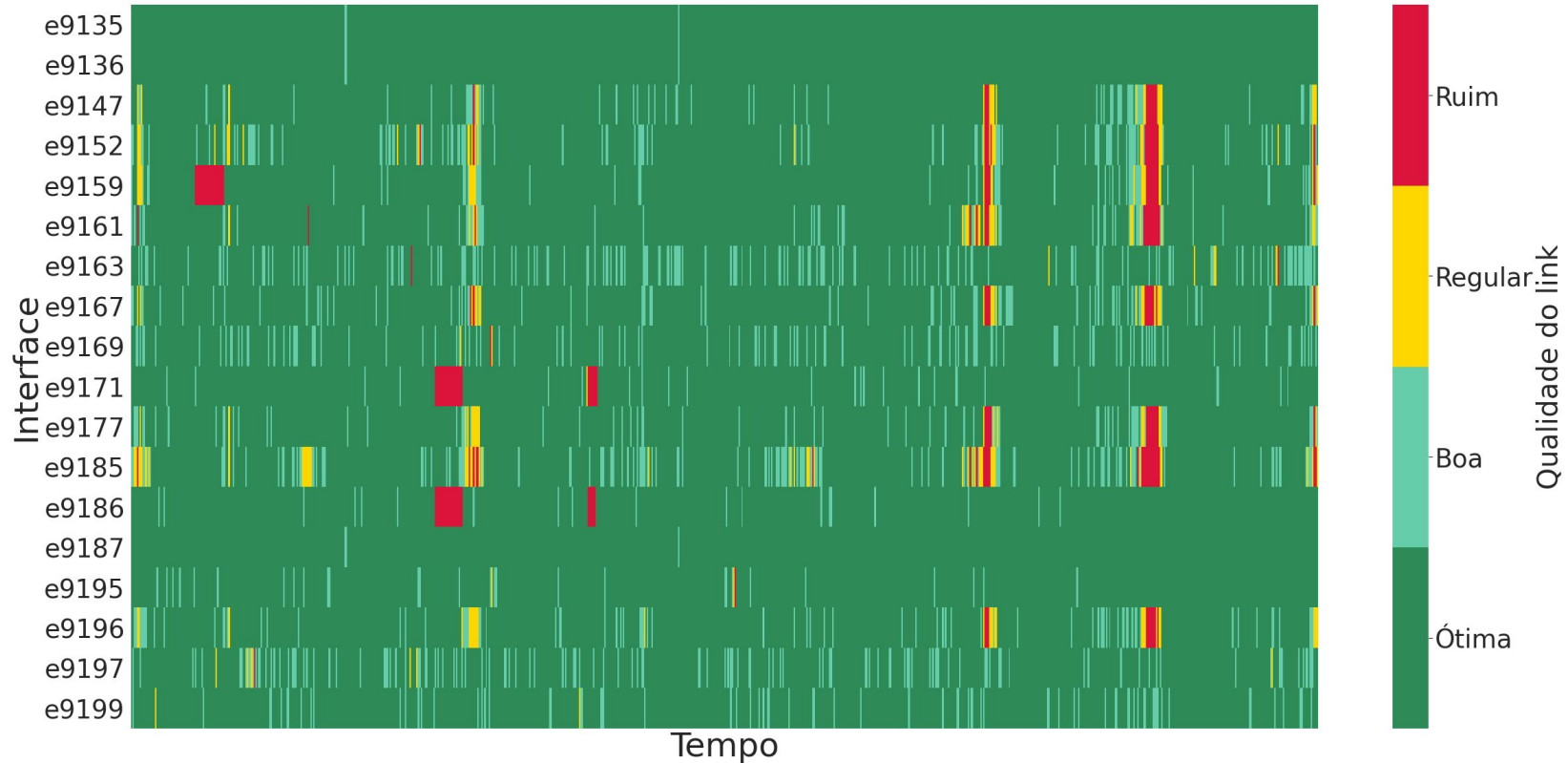
Para fins deste trabalho consideramos uma interface num estado de **falha** se a perda for classificada como **Ruim**.

Caracterizações iniciais das Falhas

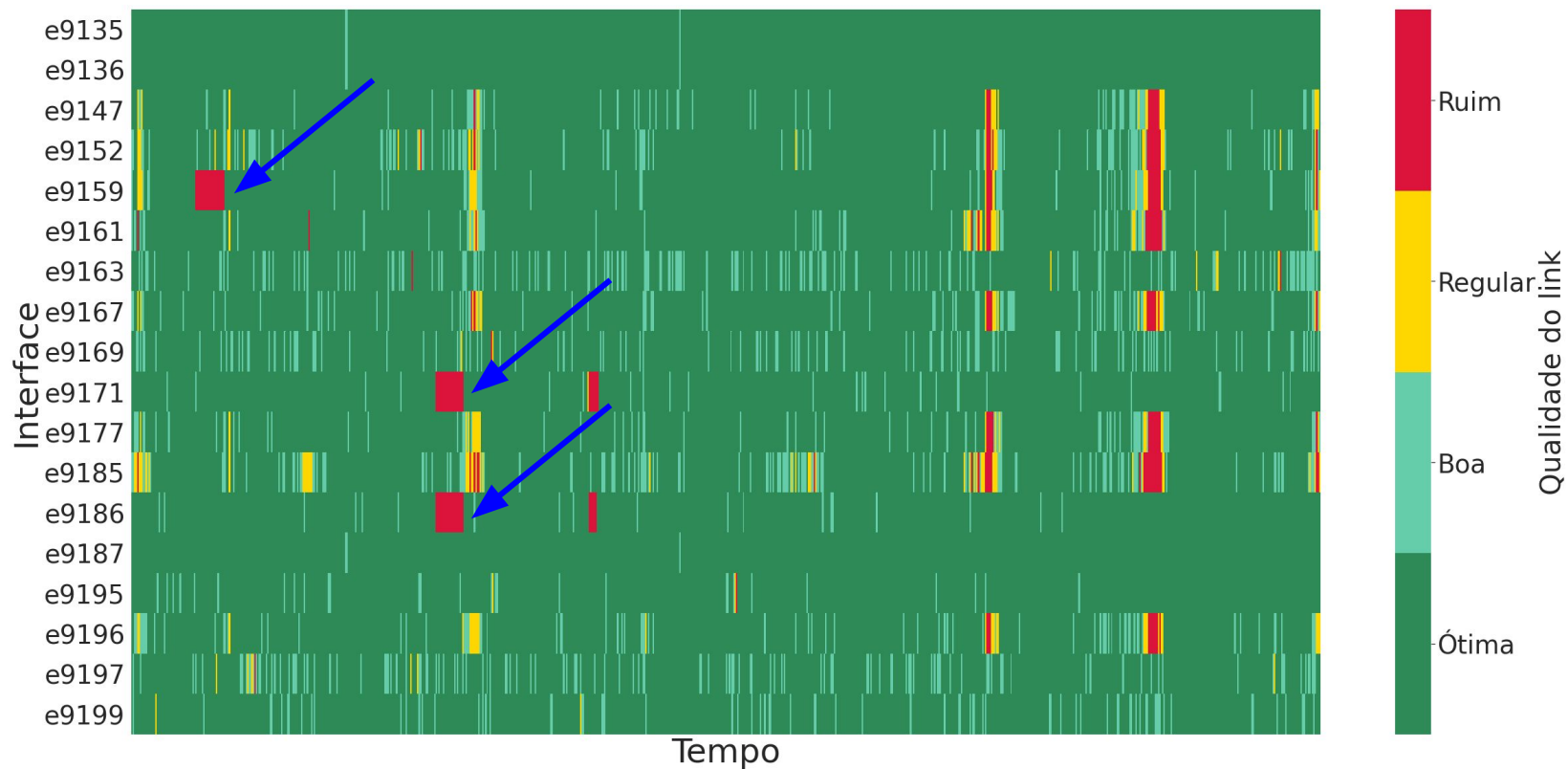
Extraímos três métricas relacionadas às falhas em cada interface:

- Duração das falhas
- Tempo entre falhas
- Número de falhas

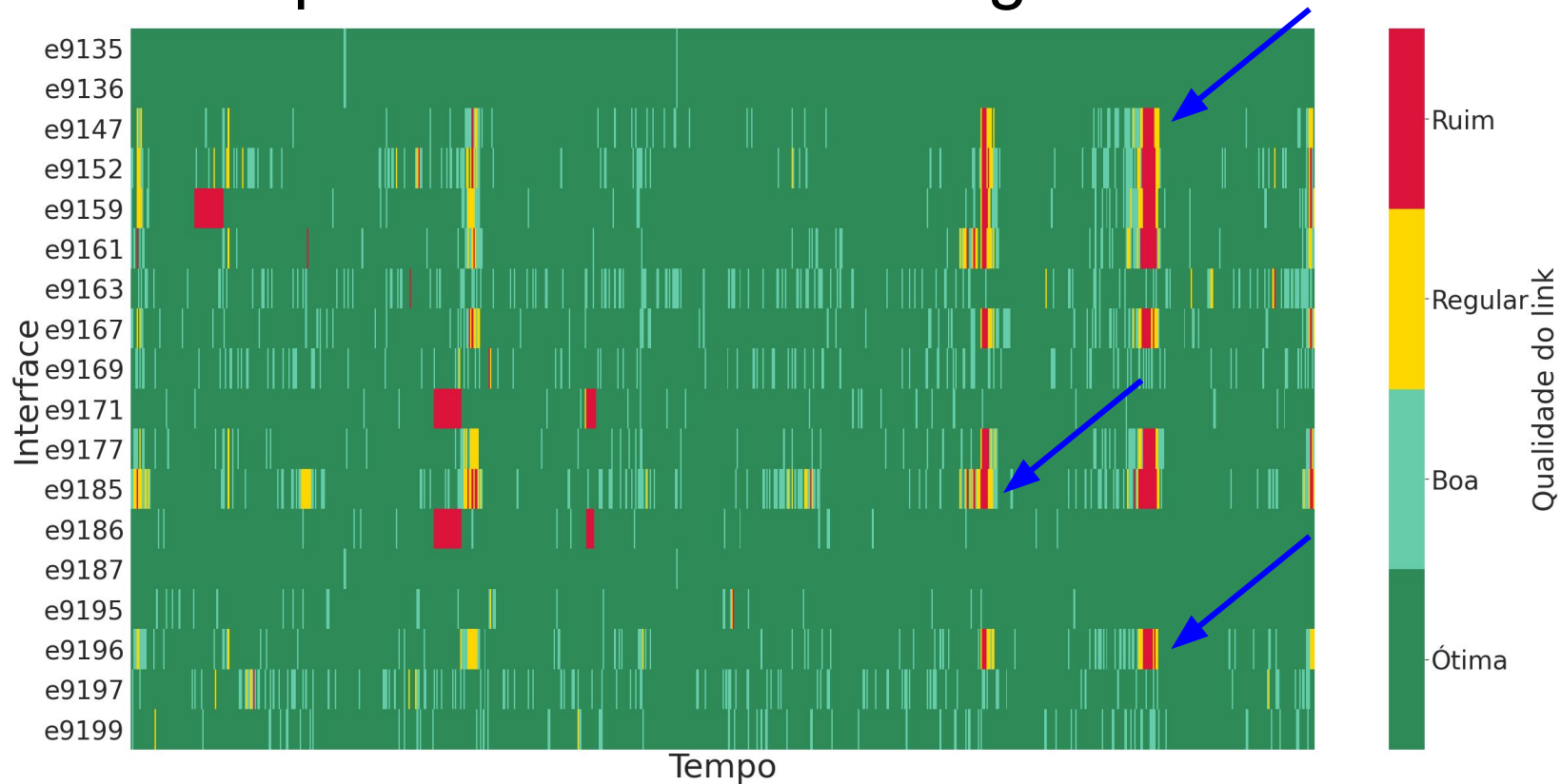
Visualização de Falhas no ES (semana 16-22/abril)



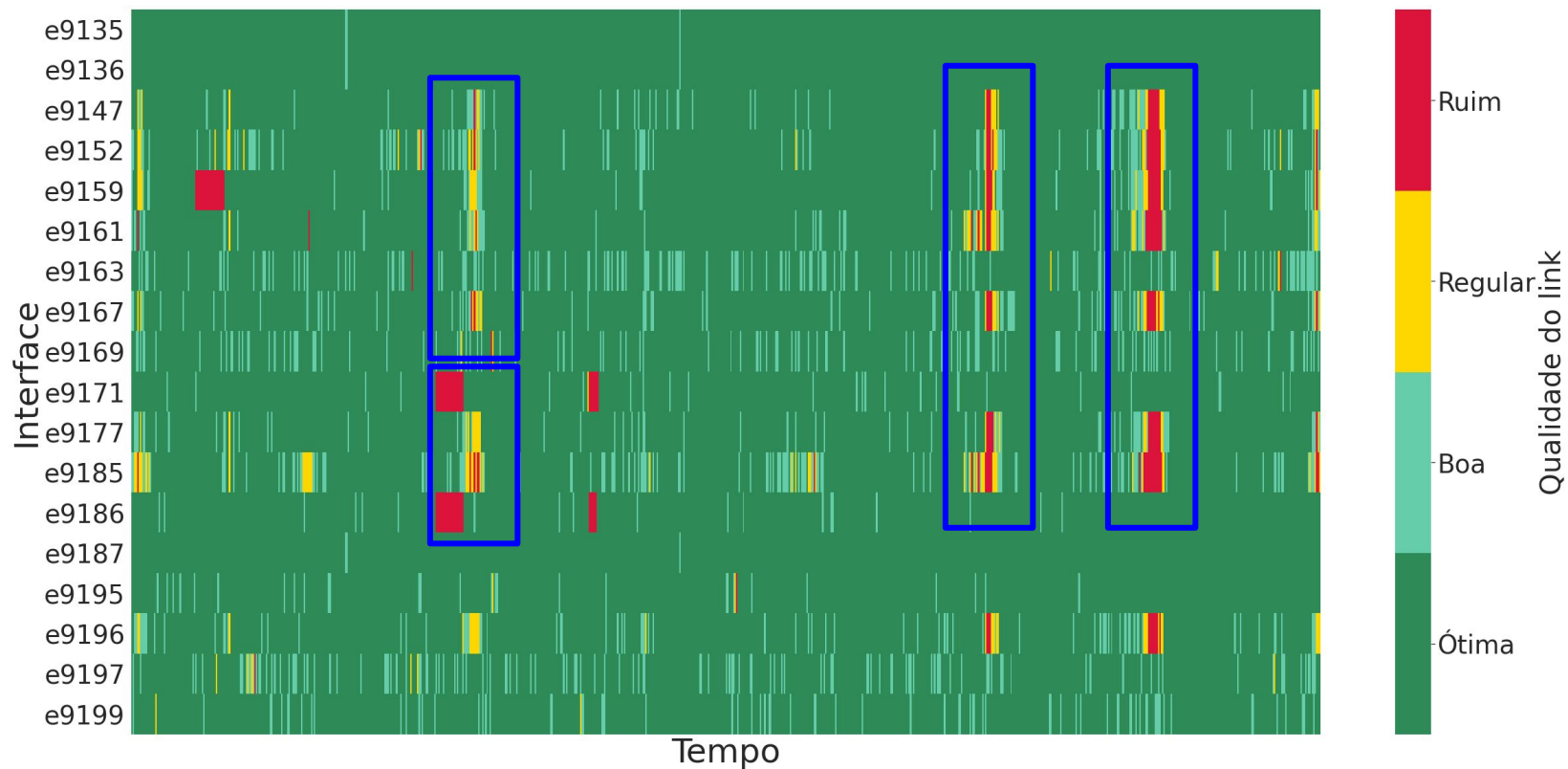
Há falhas que ocorrem repentinamente



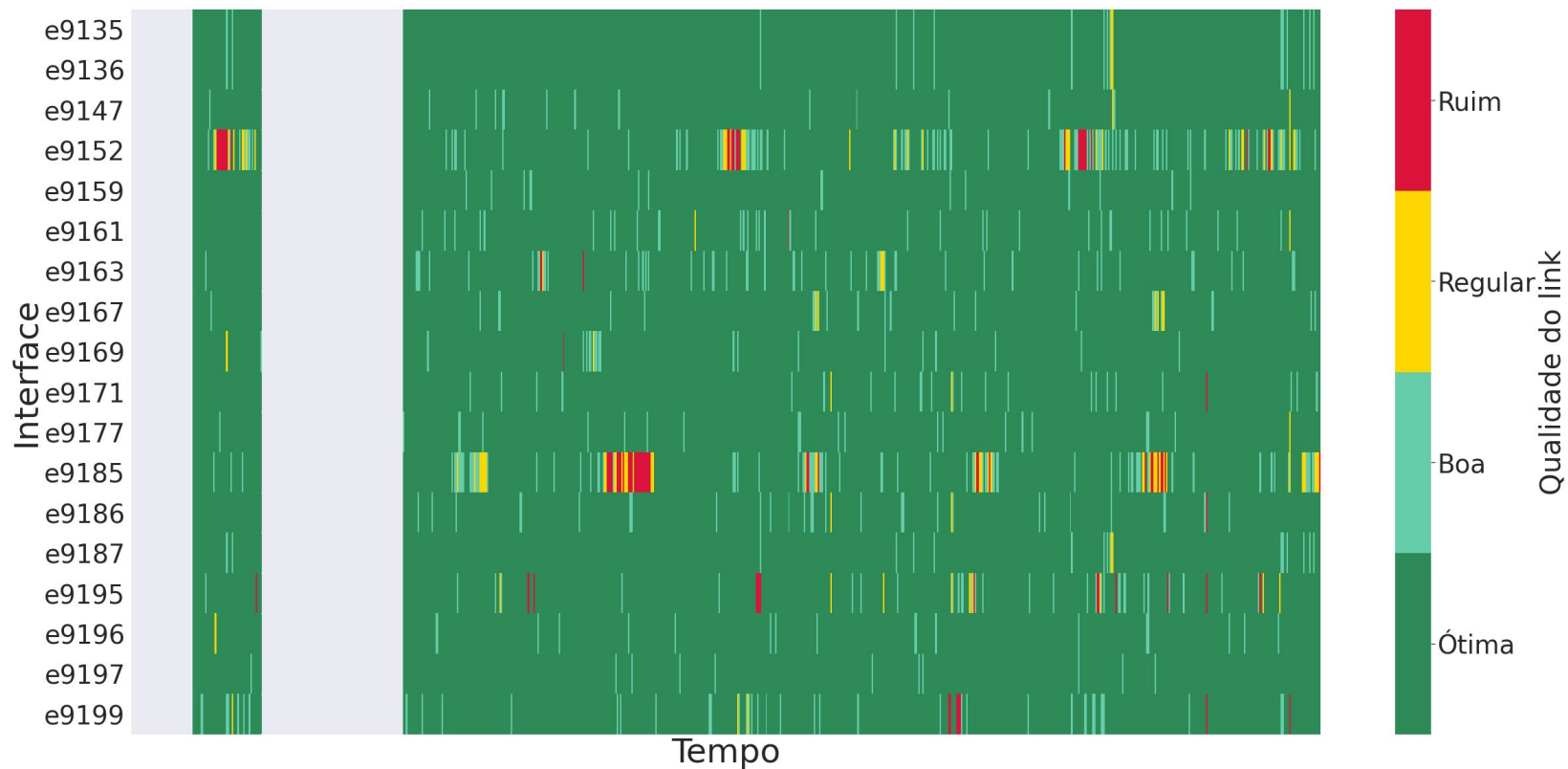
Há falhas que ocorrem de forma gradual



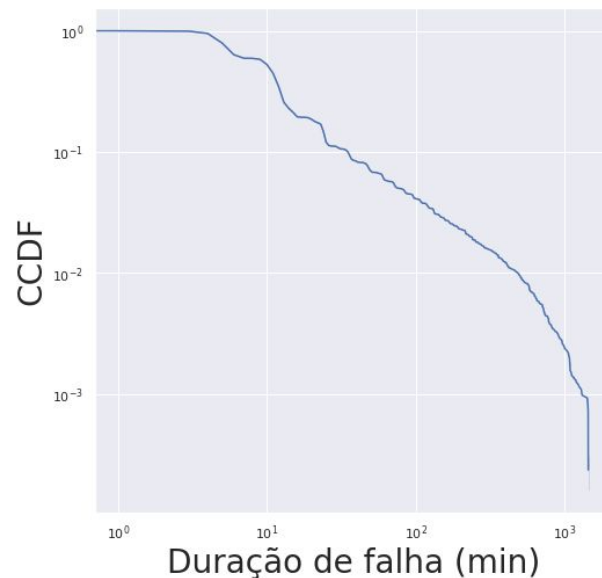
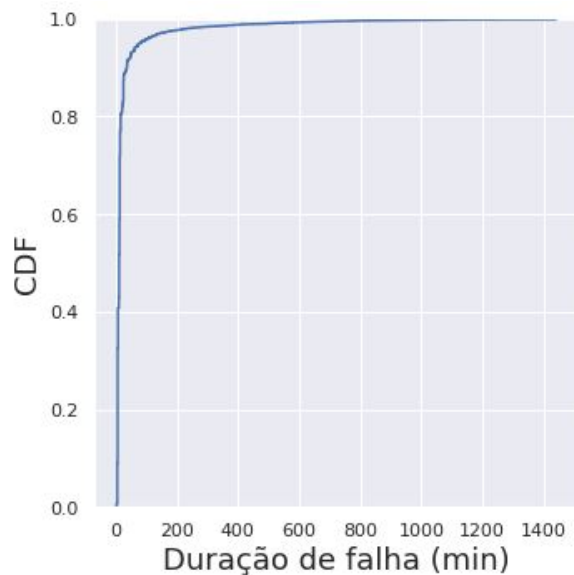
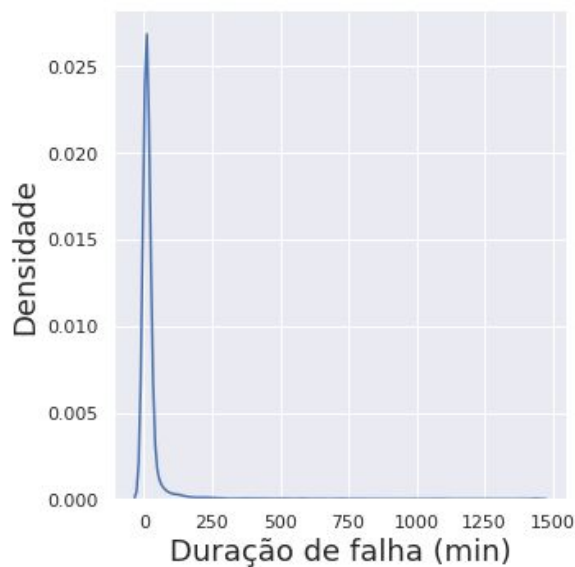
Há correlação espacial



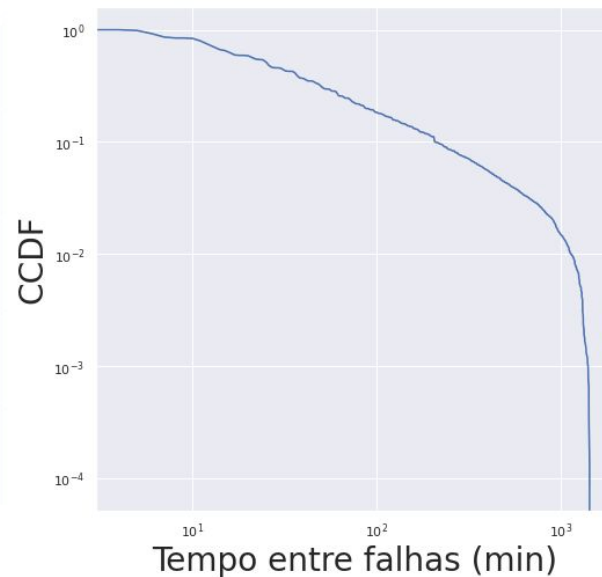
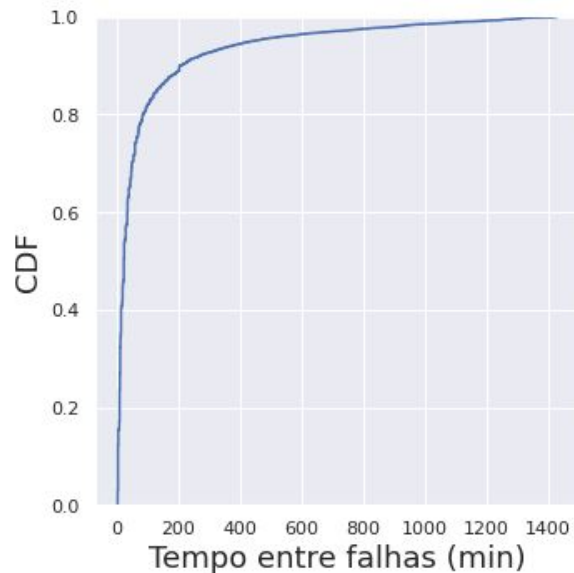
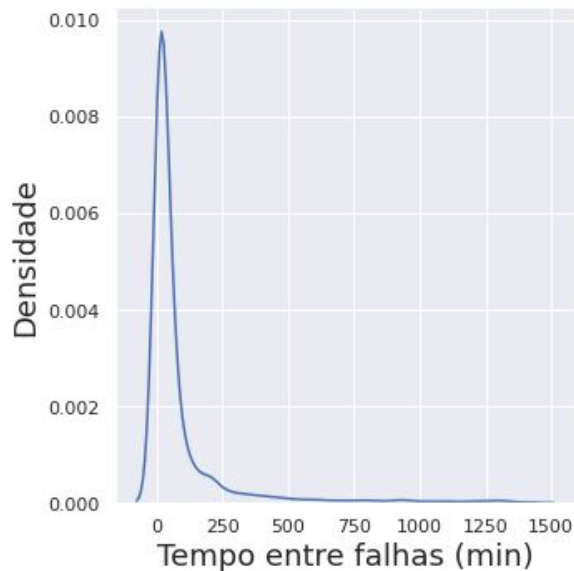
Falhas no ES na semana (10-16/julho) seguem as características



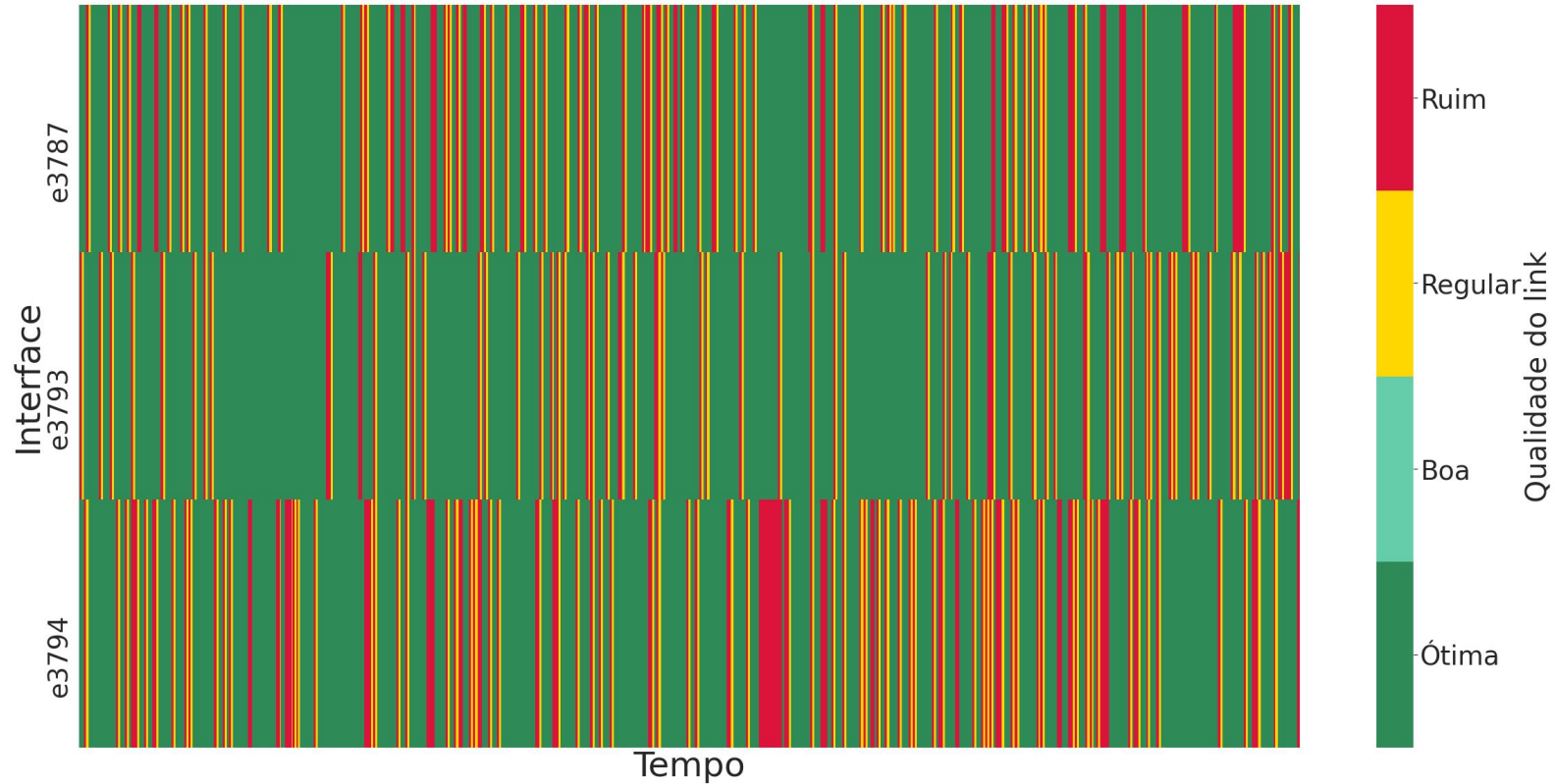
A maior parte das falhas dura pouco tempo



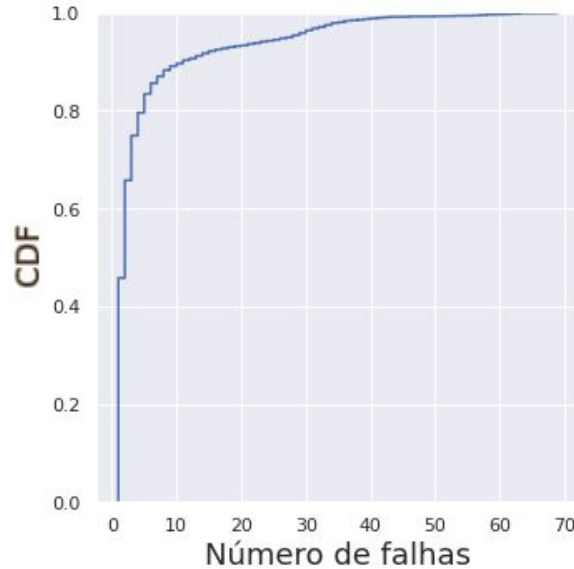
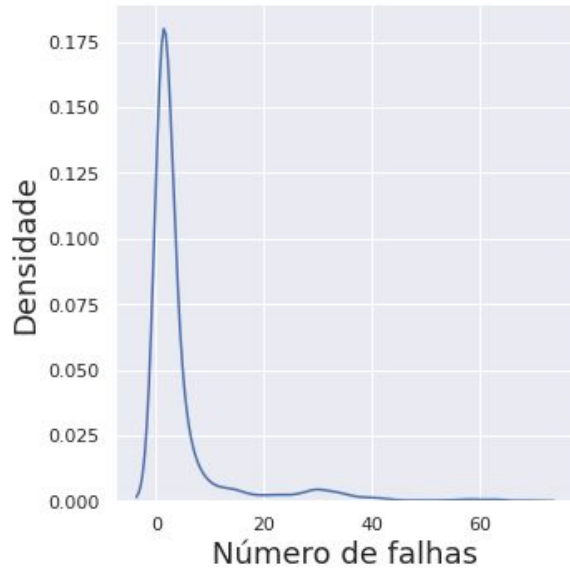
Falhas ocorrem em rajadas



Falhas ocorrem em rajadas (e.g., 3 interfaces do AM)



Muitas interfaces têm poucas falhas
Poucas Interfaces têm muitas falhas



Problema de Previsão

Considere

s_t : estado da rede no tempo t

Por agora, vamos assumir que o **estado** da rede seja dado pela fração de pacotes perdidos.

Problema de Previsão

Considere

s_t : estado da rede no tempo t

Por agora, vamos assumir que o **estado** da rede seja dado pela fração de pacotes perdidos.

Dada uma janela de tempo com w minutos

$$s_{t-(w-1)}, \dots, s_{t-2}, s_{t-1}, s_t$$

Problema de Previsão

Considere

s_t : estado da rede no tempo t

Por agora, vamos assumir que o **estado** da rede seja dado pela fração de pacotes perdidos.

Dada uma janela de tempo com w minutos

$$\underbrace{s_{t-(w-1)}, \dots, s_{t-2}, s_{t-1}}_{\text{Passado}}, \underbrace{s_t}_{\text{Presente}}$$

Problema de Previsão

Considere

s_t : estado da rede no tempo t

Por agora, vamos assumir que o **estado** da rede seja dado pela fração de pacotes perdidos.

Dada uma janela de tempo com w minutos

$$\underbrace{s_{t-(w-1)}, \dots, s_{t-2}, s_{t-1}, s_t}_{\text{Passado}} \quad \overbrace{s_t}^{\text{Presente}}$$

O que acontece no tempo $t+k$?

$$s_{t+k} \overset{?}{>} 3\%$$

Problema de Previsão

Considere

s_t : estado da rede no tempo t

Por agora, vamos assumir que o **estado** da rede seja dado pela fração de pacotes perdidos.

Dada uma janela de tempo com w minutos

$$\underbrace{s_{t-(w-1)}, \dots, s_{t-2}, s_{t-1}, s_t}_{\text{Passado}} \quad \overline{s_t}_{\text{Presente}}$$

O que acontece no tempo $t+k$?

$$\underbrace{s_{t+k} \stackrel{?}{>} 3\%}_{\text{Futuro}}$$

Problema de Previsão

Considere

s_t : estado da rede no tempo t

Por agora, vamos assumir que o **estado** da rede seja dado pela fração de pacotes perdidos.

Dada uma janela de tempo com w minutos

$s_{t-(w-1)}, \dots, s_{t-2}, s_{t-1}, s_t$

Passado Presente

O que acontece no tempo $t+k$?

$s_{t+k} > 3\%$ Futuro

Por agora:
 $w = 30$ e $k = 1$

Conjuntos de treino, validação e teste

Para o avaliar a LSTM, escolhemos a primeira semana e dividimos os dados da seguinte forma:

- **Treino:** 16/04 até 20/04
- **Validação:** 21/04
- **Teste:** 22/04

Cr terios de Avalia  o

Para avaliar o resultado calculamos a matriz de confus  o:

		VALOR PREDITO	
		0	1
VALOR REAL	0	True Negative	False Positive
	1	False Negative	True Positive

Métricas de Avaliação

Precisão $\frac{TP}{TP + FP}$

Acurácia $\frac{TP + TN}{TP + FN + TN + FP}$

Revocação $\frac{TP}{TP + FN}$

F1-score $\frac{2 * precision * recall}{precision + recall}$

Resultados Iniciais no Brasil

Accuracy: 98.8%

Precision: 97.9%

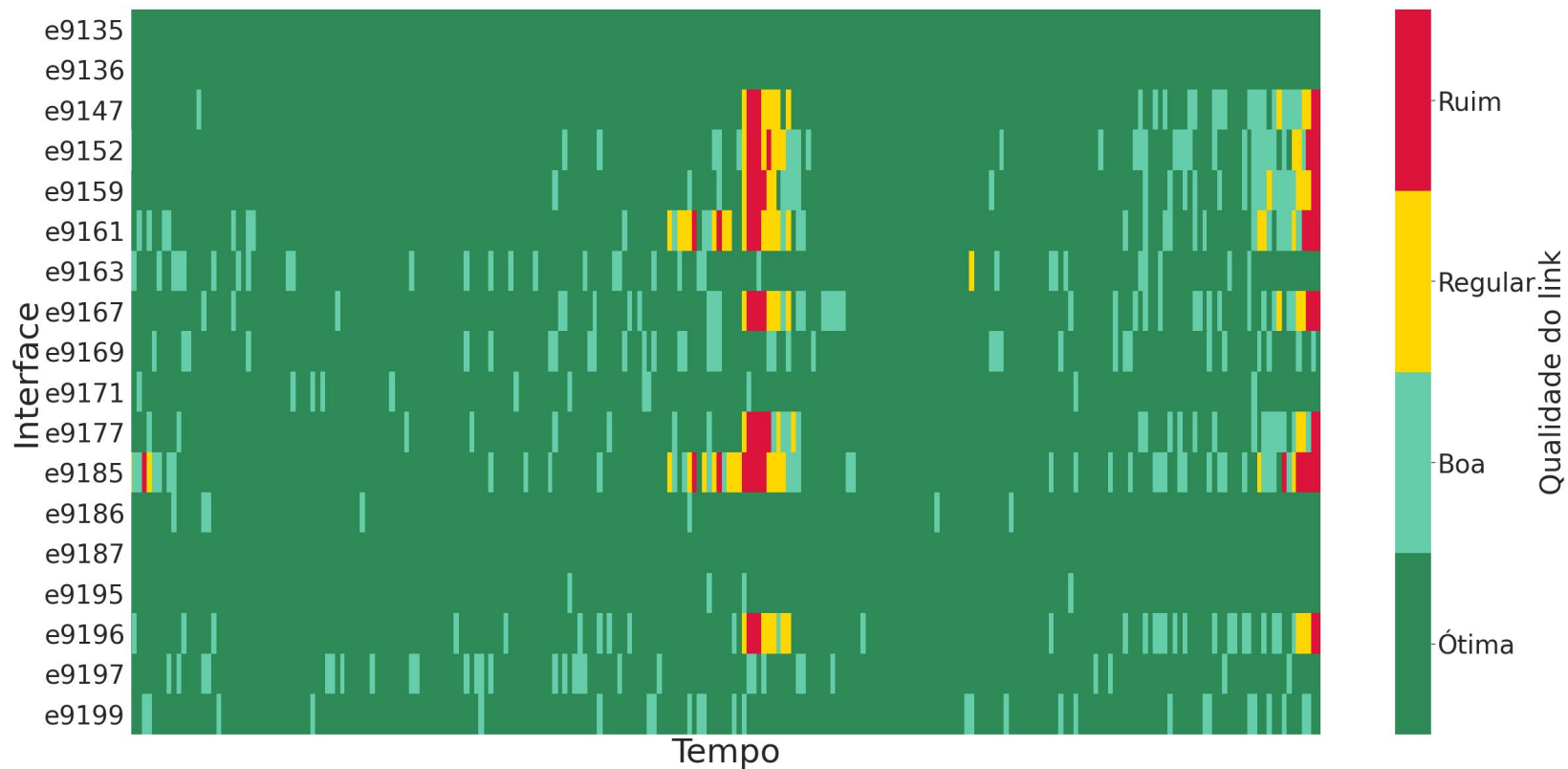
Recall: 96.7%

F1: 97.3%

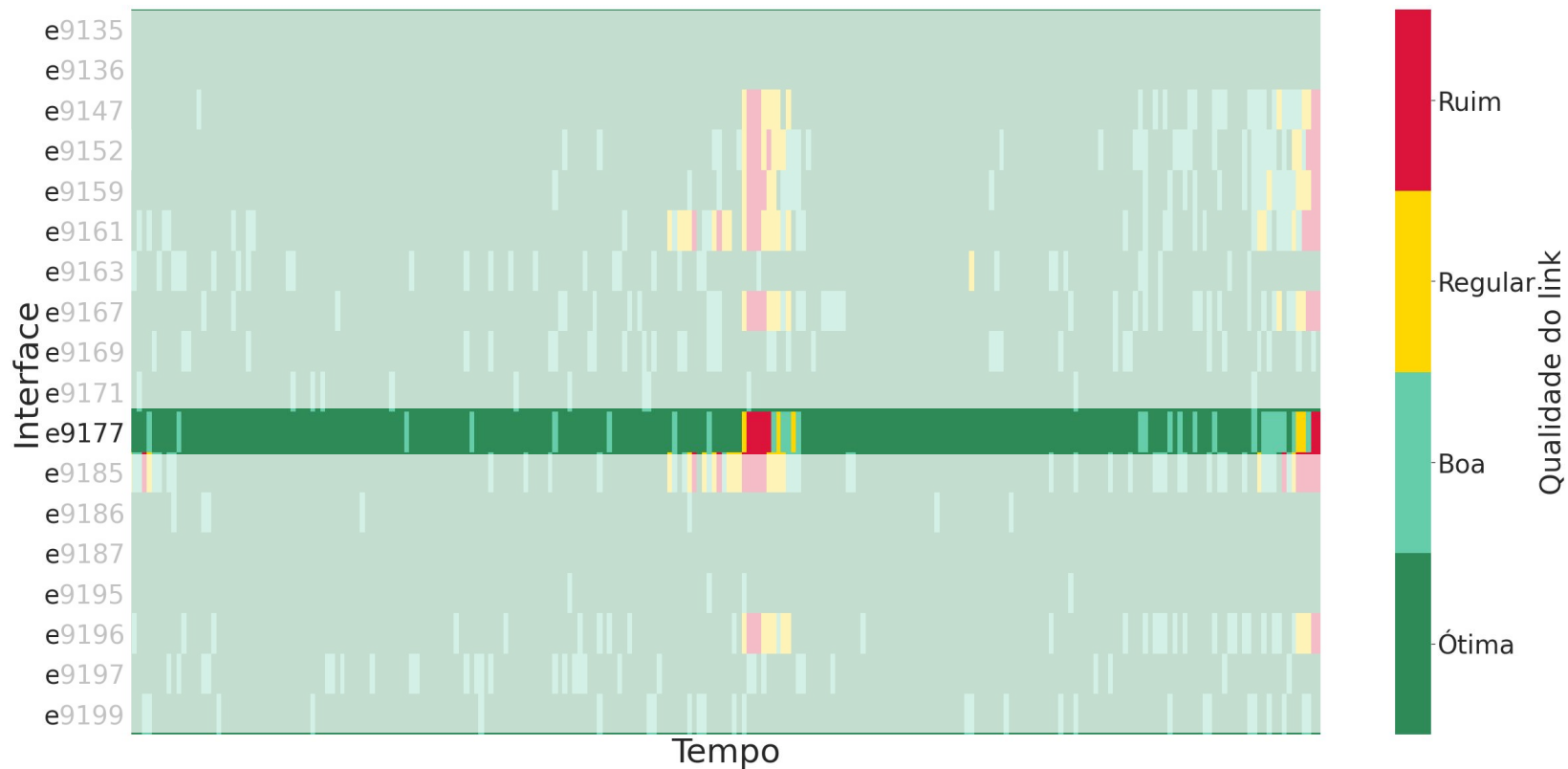
VALOR PREDITO

		VALOR PREDITO	
		0	1
VALOR REAL	0	177160	951
	1	1538	45179

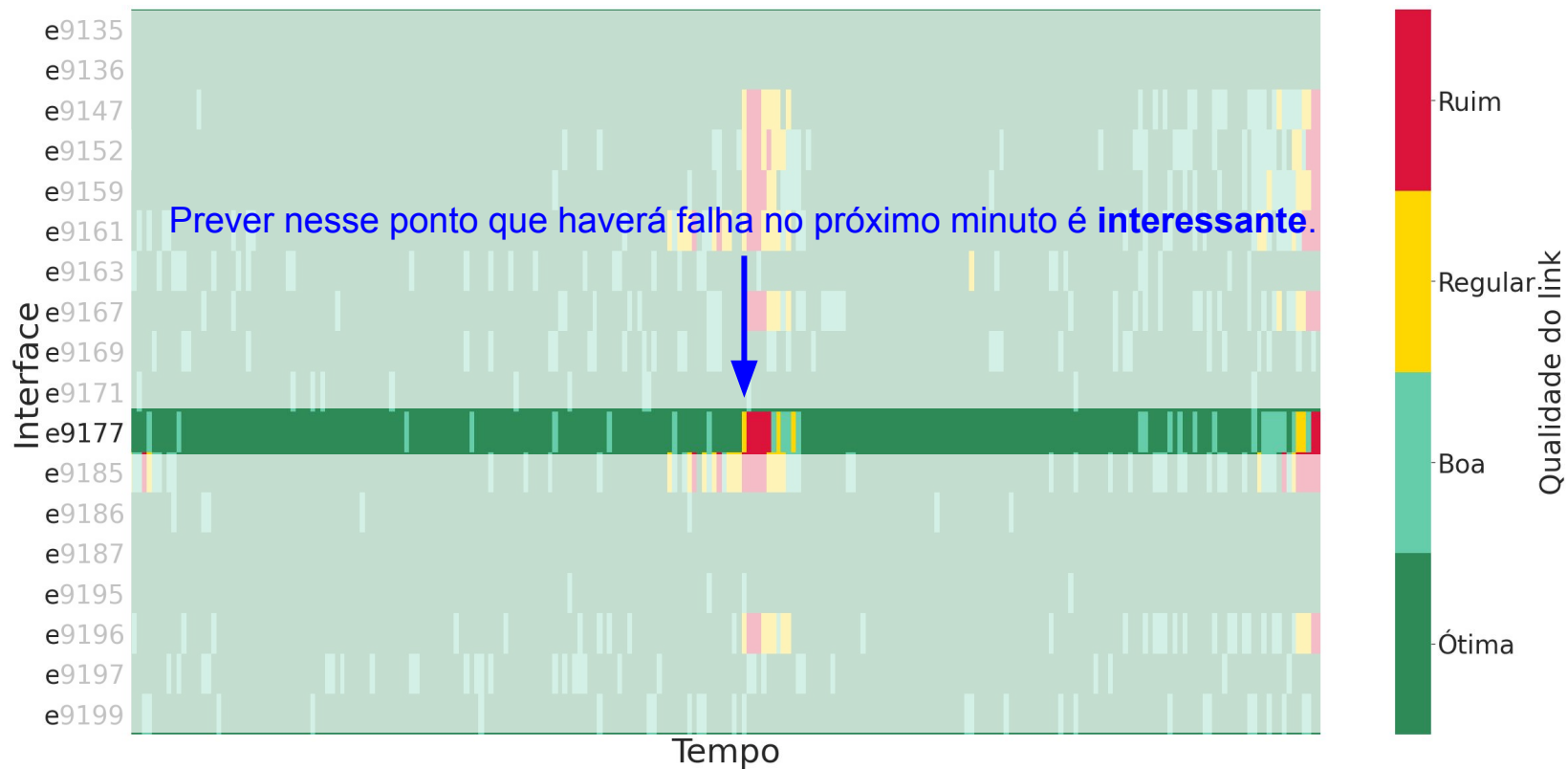
Quando é interessante prever uma falha?



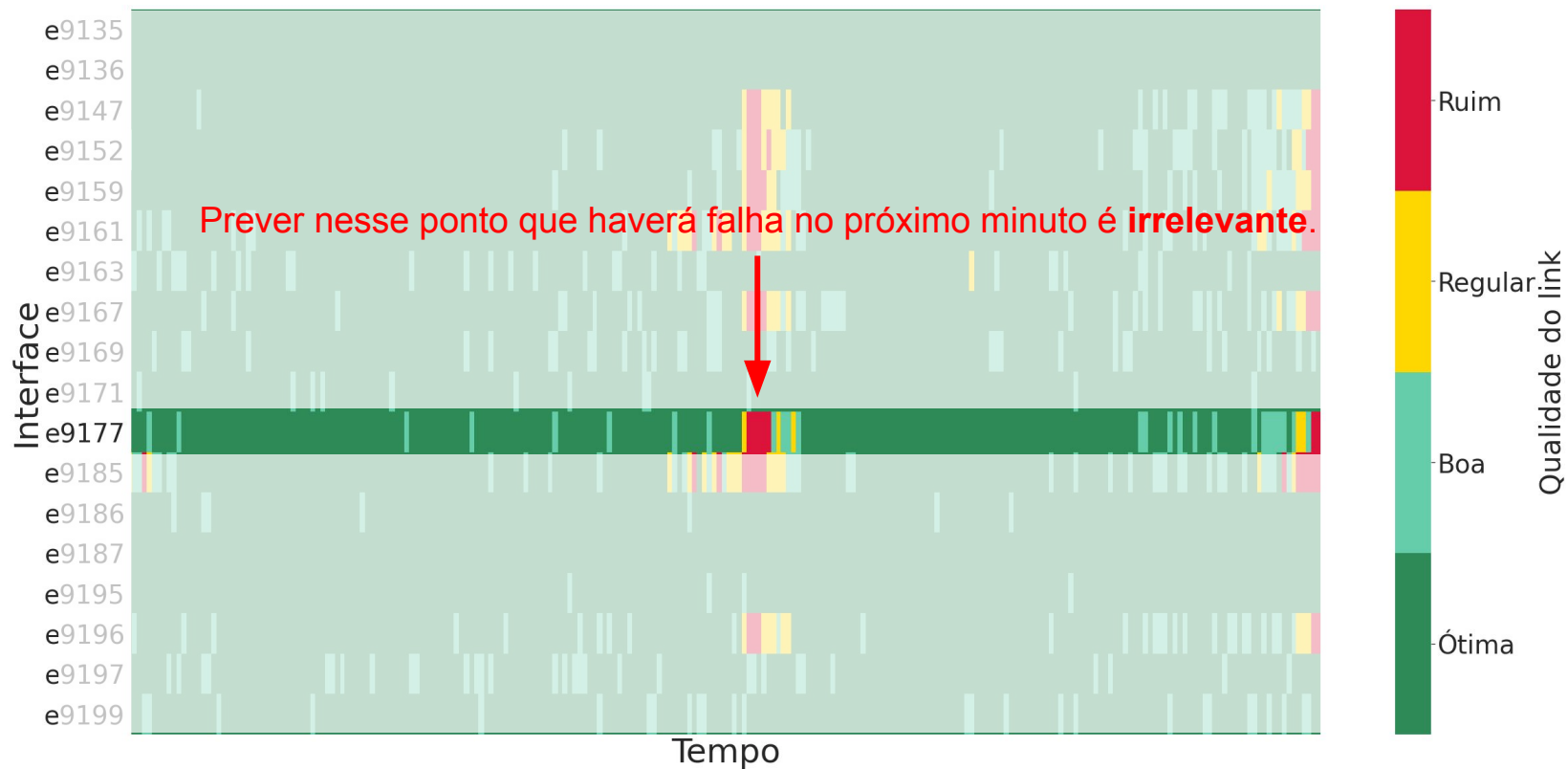
Quando é interessante prever uma falha?



Quando é interessante prever uma falha?



Quando é interessante prever uma falha?



O problema é bem mais desafiador

Accuracy: 97.8%

Precision: 24.3%

Recall: 75.1%

F1: 36.7%

VALOR PREDITO

		VALOR PREDITO	
		0	1
VALOR REAL	0	172595	3542
	1	377	1140

Considerações finais e futuros desafios

Modelagem ainda precisa ser amadurecida

- Meta é atingir uma **gerência preditiva** para os próximos 10 ou 15 minutos, considerando $k > 1$;
- **Análise da correlação espacial** a partir das topologias relacionando com as operadoras e seus respectivos enlaces (domínio adm.).

Planejamos considerar dados referentes à outras fontes de falhas

- Entender as **categorias/causas das “faltas”** ($p_loss > 3\%$) : quedas de enlaces, energia, equipamento ou configuração ;
 - A localização da falha (pop, provedor, entidade)

Dashboard, alarmes preditivos e modelo de aprendizado, integrados ao viaipe